

HEALTH CARE COST ANALYSIS AND PREDICTION

Submitted in partial fulfillment of the requirements for
the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

AKSHAYA DANTHA (Reg. No. 37110029)

AMRAPALI BHOWMIK(Reg. No. 37110037)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI
SALAI, CHENNAI – 600 119

MARCH - 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this project report is the bonafide work of **AKSHAYA DANTHA** (Reg. No. 37110029) and **AMRAPALI BHOWMIK** (Reg. No.36110776) who carried out the project entitled "**ONLINE CRIME REPORTING SYSTEM**" under my supervision from **November 2021** to **March 2021**.

Internal Guide
Dr. P. JEYANTHI.,

PROFESSOR , DEPT OF IT

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I **AKSHAYA DANTHA** hereby declare that the Project Report entitled “**HEALTH CARE COST ANALYSIS AND PREDICTION**” is done by me under the guidance of **Mrs. P. JEYANTHI.**, Department of IT at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala, M.E., Ph.D., Dean**, School of Computing, **Dr. S. Vigneswari, M.E., Ph.D., and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Mrs. P. JEYANTHI, Professor**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

This project studies the temporal consistency of health care expenditures in a large state Medicaid program. Predictive machine learning models were used to forecast the expenditures, especially for the high-cost, high-need (HCHN) patients. We systematically tests temporal correlation of patient-level health care expenditures in both the short and long terms. The results suggest that medical expenditures are significantly correlated over multiple periods. Our work demonstrates a prevalent and strong temporal correlation and shows promise for predicting future health care expenditures using machine learning. Temporal correlation is stronger in HCHN patients and their expenditures can be better predicted. Including more past periods is beneficial for better predictive performance. This study shows that there is significant temporal correlation in health care expenditures. Machine learning models can help to accurately forecast the expenditures. These results could advance the field toward precise preventive care to lower overall health care costs and deliver care more efficiently.

TABLE OF CONTENTS

ABSTRACT	v
LIST OF FIGURES	viii

CHAPTER No.	TITLE	PAGE No.
1.	INTRODUCTION	9
2.	LITERATURE SURVEY	11
3.	AIM AND SCOPE OF THE PRESENT INVESTIGATION	10
	3.1 AIM	11
	3.2 SCOPE OF INVESTIGATION	12
4.	EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHMS USED	14
	4.1 DESCRIPTION OF DIAGRAM	14
	4.2 ARCHITECTURAL DESIGN	16
	4.3 ALGORITHMS	17

5.	SUMMARY AND CONCLUSIONS	19
6.	PLAGIARISM REPORT	20
7.	PAPER WORK	27
8.	SOURCE CODE	22

CHAPTER 1

INTRODUCTION

Health care is one of the largest components of the global economy. According to the World Bank, in 2019, health care expenditures accounted for 9.95% of the world's total gross domestic product (GDP). This amount is expected to increase over the next several years, because of the expansion of insurance coverage under the Affordable Care Act. In addition, resources, in terms of expenditures, are disproportionately consumed by a relatively small proportion of the health care utilizing population. As a result, this group of health care utilizers has been termed high-cost, high-need (HCHN) patients due to its disproportionate spending concentration and highly prevalent comorbid chronic condition profile. Stakeholders have argued the need for higher efficiency health care, especially for HCHN patients. For example, managed care organizations and health plans (MCOs) have been deployed and funded under capitation payment systems to incentivize health care providers to deliver more cost-effective services.

Information technology provides a new, promising way to approach a wide-range of health care problems, especially in the "Big Data" era. Health care utilization routinely generates vast amounts of data from sources ranging from electronic medical records, insurance claims, vital signs, and patient-reported outcomes. Moreover, predicting health outcomes using data modeling approaches is an emerging field that can reveal important insights into factors associated with disproportionate spending patterns. Specifically, if we are able to forecast expenditures at the patient-level with good accuracy, we could improve targeted care by anticipating health care needs of HCHN patients. Moreover, predictive modeling can improve our understanding of causal pathways leading to our understanding of expensive events which informs system-level strategies for prevention. Indeed, prevention one of the most effective ways to lower health care expenditures while delivering better quality of care.

HCHN patients have shown a high degree of persistence in medical expenditures reflected in administrative claims data, although some recent studies have concluded that the high utilization may be temporary and not persistent. The purpose of this study is to better understand these temporal patterns and to apply machine learning-based models to predict expenditures. Additionally, we would like to determine if practical timeframes (e.g. 1 month, 6 months) are feasible choices for predicting expenditures.

The health care executives have tremendous effect on the availability, accessibility and quality of health. The dramatic changes in industry due to new technologies along with growth in population have significantly increased the demand for quality health care services and the responsibility is majorly taken up by the administrators/managers of such organizations to implement cost containment strategies. To continue meeting community health care needs in this new delivery and payment environment, hospital and health system leaders will need to think and act strategically about managing cost. Strategic cost transformation will be required, and to address this issue, the study was undertaken to know what city-based hospitals are doing to maintain the norm of quality health care in optimum cost.

CHAPTER 2

LITERATURE SURVEY

Healthcare Management System is achieved with the advent of Predictive Analytics. While base line of a Predictive Analytic Tool is the same, the tool comes in various versions. Several developers aim at developing a tool that is efficient than all other tools developed so far. The following papers have implemented some of the most prevalent tools.

Data mining is defined as the analysis of often large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Data mining in the field of Healthcare is a gradually developing methodology that is used to identify a legitimate, useful relationships and patterns in the data that are too complex and subtle for human to make out. The previously unknown patterns and relationships in a database can be easily identified with a Data mining algorithms like classification, association, clustering, etc and with this we can build a Predictive Model. To identify the unknown patterns, we use the life cycle model of data mining. It includes all the process from data selection to data exploitation.

Data mining is beneficial for all those who are involved in the healthcare including Healthcare Insurers, Healthcare organizations, Doctors and patients. For healthcare insurers, data mining can help identify Fraud and Abuse in the insurance. For Healthcare organizations, it supports Hospital Management system, Patient Tracking system and thereby forming a strong Customer relationship management. For doctors and physicians, it can help predict the LOS, mortality rate, risk factors, and re-admission rate. With all the aforesaid, a doctor can choose upon the effective treatment available for the patient. And finally, a patient can get a quality treatment at affordable cost. According to a survey, 87% of the deaths in the US hospital have been prevented with the Predictive tool.

CHAPTER 3

AIM & SCOPE OF PRESENT INVESTIGATION

3.1 AIM

Predictive modeling include risk adjustment models that are linear in design and form the basis of many capitation payment systems .This models suffer from a number of limitations including use of variables with limited predictive accuracy, specific patient populations or type of care, and population-level models that offer limited information at the patient-level .The authors in analyzed the temporal utilization pattern of high utilizers in a large public state insurance program .Studies that attempt to predict HCHN status using patient-level expenditures are lacking. This study aims to address this gap in the literature.

The importance of electronic data has increased manifold over the last decade. With a consistent improvement in computing capabilities, substantial amounts of electronic data are being collected that can provide answers to critical research questions. In the high-impact field of healthcare informatics also, the utility of gathering electronic information has been realised. A part of this endeavour is a unique community health data system called Arizona HealthQuery (AzHQ), housed in the Centre for Health Information and Research (CHiR) at Arizona State University. AzHQ contains comprehensive health records of over eight million patients linked across time from the state of Arizona, USA. AZHQ offers the opportunity for research that can influence the community and deliver actionable results to researchers and policy makers. The surge in data accumulation was accompanied by the development of improved methods of obtaining knowledge from the data. Data mining and machine learning methods have contributed immensely to this important task. These techniques have been successfully applied to many domains but applications to healthcare are still relatively rare (Anderson et al., 2004; Cios and Moore, 2002; Li et al., 2005). Using such techniques, this paper describes a potentially beneficial approach to healthcare, focusing on the difficult and important problem of predicting the patients who are likely to be high-risk. One of the most important issues dogging the healthcare system is that approximately two-thirds of healthcare costs in the USA are accounted for by 10% of the patients.

3.2 SCOPE OF INVESTIGATION

When the project is about to be funded, there should be a set of defined deliveries and objectives for the project. There can be a high level-scope statement prepared at this level.

This high-level scope statement can be taken from the initial documents such as SOW. In addition to the SOW, you need to use any other document or information in order to further define the project scope at this level.

In case, if you feel that you do not have enough information to come up with a high-level scope statement, you should then work closely with the client in order to gather necessary information.

Project objectives can be used for defining the project scope. As a matter of fact, there should be one or more deliveries addressing each project objective in the project. By looking at the deliverables, you can actually gauge the project scope.

Once you get to know the main deliverables of the project, start asking questions about the other processes and different aspects of the project.

First identifying and clearly defining the out of scope also helps you to understand the scope of a project. When you go on defining the out of scope, you will automatically get an idea of the real project scope. In order to follow this method, you need to have a defined scope up to a certain level.

Whenever you identify an item for the scope or out-of-scope, make sure you document it then and there. Later, you can revisit these items and elaborate more on those.

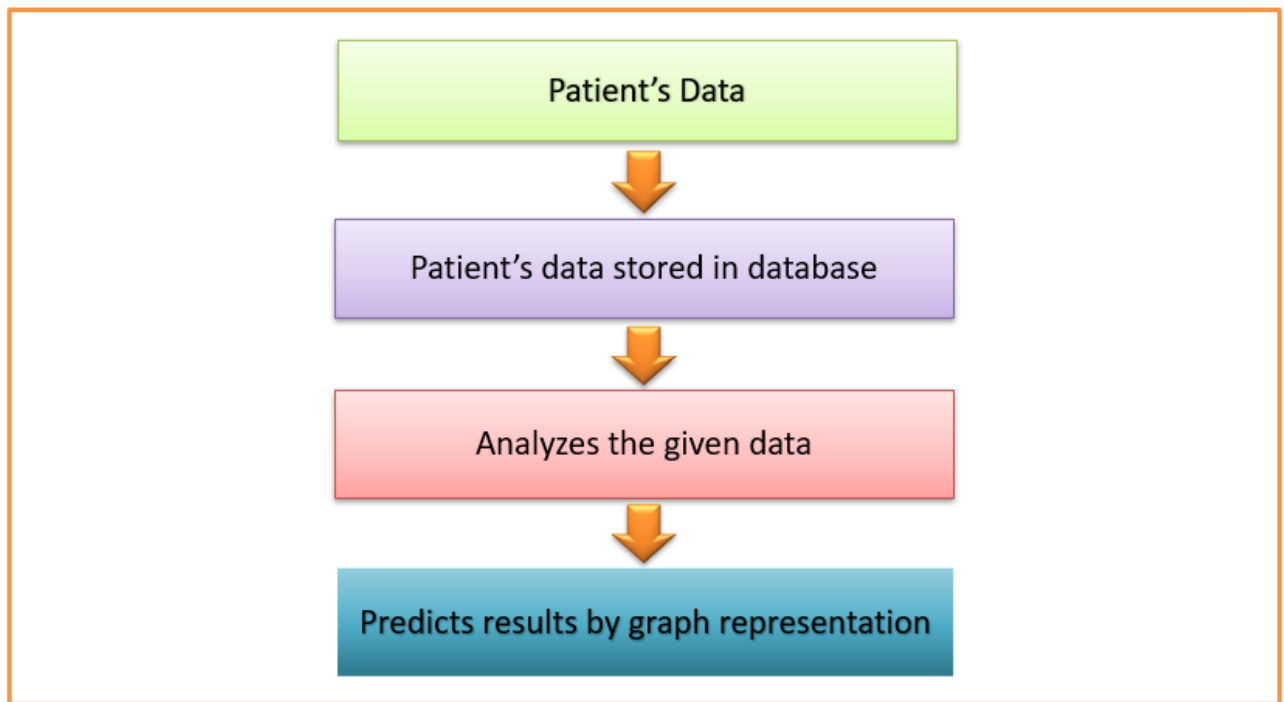
Once you have successfully defined the scope of the project, you need to get the sign-off from the related and applicable parties. Without proper sign-off for the scope, the next phases of the project, i.e., requirements gathering, might have issues in executing.

CHAPTER 4

EXPERIMENTAL OR MATERIALS AND METHODS & ALGORITHMS

4.1 DESCRIPTION OF DIAGRAM

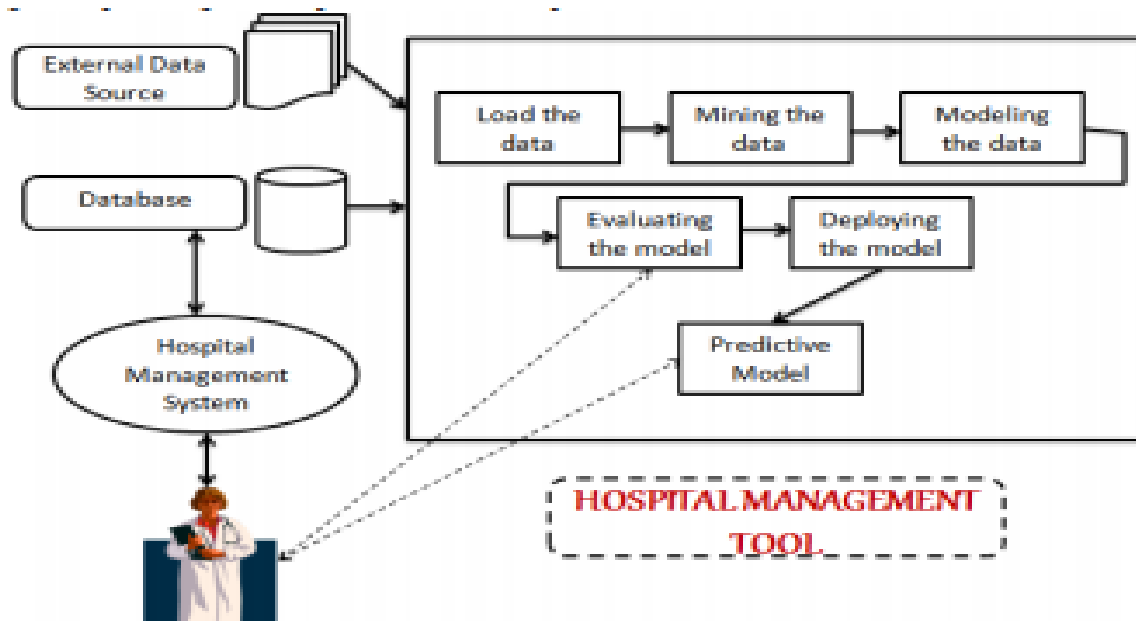
It takes raw data given, later it analyzes the given data and stores it in data base and further it predicts the by graph representation. It is a way to examine both the costs and health outcomes of one or more interventions. It compares an intervention to another intervention (or the status quo) by estimating how much it costs to gain a unit of a health outcome, like a life year gained or a death prevented. CEA provides information on health and cost impacts of an intervention compared to an alternative intervention (or the status quo). If the net costs of an intervention are positive (which means a more effective intervention is more costly), the results are presented as a cost-effectiveness ratio. A cost-effectiveness ratio is the net cost divided by changes in health outcomes. Examples include cost per case of disease prevented or cost per death averted. However, if the net costs are negative (which means a more effective intervention is less costly), the results are reported.



Studies of costs and related economic implications comprise a major group of methods used in HTA. These studies can involve attributes of either or both of primary data collection and integrative methods. That is, cost data can be collected, for example, as part of RCTs and other clinical studies as well as administrative (“claims”) databases used in health care payment. Cost data from one or more such sources often are combined with data from primary clinical studies, epidemiological studies, and other sources to conduct cost-effectiveness analyses and other analyses that involve weighing health and economic impacts of health technology.

Interest in cost analyses has accompanied concerns about rising health care costs, pressures on health care policymakers to allocate resources, and the need for health product makers and other technology advocates to demonstrate the economic benefits of their technologies. This interest is reflected in a great increase in the number of reports of cost analyses in the literature and further refinement of these methods.

4.1 ARCHITECTURAL DESIGN



Supervised learning problems can be further grouped into

- Regression problems and
- Classification problems

In classification, learning algorithms takes the input data and map the output to a discrete output like True or False In regression, learning algorithms maps the input data to continuous output like weight, cost, etc.

In this project I will apply regression techniques of supervised learning to predict the medical insurance costs.

Dataset consists of records. Each record contains the below data for specific person.

- age – Age of the person
- sex – Sex of the person
- bmi – Body Mass Index(BMI) of the person
- children – Number of children for the person
- smoker – Smoking status of the person
- region – Region of the person in US
- charges – Medical Insurance costs per year for the person

The health care industry figures among the fast-growing industries in India. It comprises hospitals, medical equipment manufacturers, health insurers and educational institutions. The health care services in India are provided through public hospitals, corporate hospitals with super specialty facilities, private and charity hospitals, rehabilitation centres, small and medium nursing homes, private clinics and diagnostic centres.

While the changing health care situation in the country has many affirmative outcomes, it also gives rise to the challenge of bringing health care services within the reach of the poor. The increasing cost of health care, owing to a substantial up-tick in quality, poses a challenge to the providers of these services. The critical success factors in the Indian health care industry are latest medical technology, information technology, rural penetration, service orientation, industry support and transformation, which are all cost intensive, on one hand, and are also essential so as to continuously provide quality health care services to the community, on the other hand. The hospital management has a responsibility to the community—to provide health care services that the community needs, at an acceptable level of quality and at the least possible cost. Cost finding and analysis can help departmental managers, hospital administrators and policy-makers to determine how well their institutions meet these public needs (Shepard et al. 2000). The hospitals need to be run as financial viable organizations.

4.3ALGORITHMS

Algorithms are used when we have unlabeled data and are trying to group together similar data points based off of features. This is mainly used to explore the data and find some structure within.

Machine learning is a method of data analysis which sends instructions (programmable code) to computers so that they can learn from data. Then, based on the learned data, they provide us the predicted results/patterns. With the help of Machine Learning, we can develop intelligent systems that are capable of taking decisions on an autonomous basis.

CHAPTER 5

RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

5.1 PERFORMANCE ANALYSIS

To understand correlation between the input or the independent variables, I classified the ages and BMI into specific categories. The above analysis shows the stats like number of male and female, smokers and non-smokers, etc. from the input data.

Taking the BMI ranges from Medline Plus, I categorized the input samples into below-weight, normal-weight, over-weight and obese.

1. Splitting the data set into training and testing subsets helps to assess the performance of the model over an independent data set. Typically, we train the model using training data subset and then evaluate the model's performance using the testing data subset, which is independent of the training data subset.
2. Splitting the data set also helps in having a check on model's overfitting.

As part of evaluation, I considered multiple regression algorithms like decision trees, Support Vector Machines for regression, etc. Based on the metrics, I choose to use decision tree technique for this project

The goal of decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

5.2 RESULTS AND DISCUSSION

A cost-benefit analysis is the process used to measure the benefits of a decision or taking action minus the costs associated with taking that action. Ideally, these decisions would be made based on comparing the potential health benefits, harms, and costs of each alternative. Doctors, patients, hospitals, health systems, third party payors, and other

health care decision-makers all face challenges in gathering the information needed for optimal health care decision-making.

When it comes to costs, when hard choices have to be made and limited resources are available, things can be confusing. There is a substantial body of literature on economic evaluation in health care, including many studies by health economists, clinical researchers, and multidisciplinary teams. Unfortunately, these studies vary in methodology and quality.

An understanding of *economic evaluation* is fundamental to knowing how choices should be made. Economic evaluation is a process that provides a systematic way to identify relevant alternatives and minimizes the chance of an important option being missed. It allows evaluation from differing viewpoints

CHAPTER 6

SUMMARY AND CONCLUSIONS

We systematically tests temporal correlation of patient-level health care expenditures in both the short and long terms. The results suggest that medical expenditures are significantly correlated over multiple periods. Our work demonstrates a prevalent and strong temporal correlation and shows promise for predicting future health care expenditures using machine learning. Temporal correlation is stronger in HCHN patients and their expenditures can be better predicted. Including more past periods is beneficial for better predictive performance.

REFERENCES

- [1] Precision Medicine Initiative (NIH). <https://www.nih.gov/precision-medicine-initiative-cohort-program>.
- [2] Lyman, Gary H., and Harold L. Moses. "Biomarker Tests for Molecularly Targeted Therapies--the Key to Unlocking Precision Medicine." *The New England journal of medicine* 375.1 (2016): 4-6.
- [3] Collins, Francis S., and Harold Varmus. "A new initiative on precision medicine." *New England Journal of Medicine* 372.9 (2015): 793-795.
- [4] Xu, Rong, Li Li, and QuanQiu Wang. "dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text." *BMC bioinformatics* 15.1 (2014): 105.
- [5] Wang, Bo, et al. "Similarity network fusion for aggregating data types on genomic scale." *Nature methods* 11.3 (2014): 333.
- [6] Tatonetti, Nicholas P., et al. "Data-driven prediction of drug effects and interactions." *Science translational medicine* 4.125 (2012): 125ra31-125ra31.
- [7] Libbrecht, Maxwell W., and William Stafford Noble. "Machine learning applications in genetics and genomics." *Nature Reviews Genetics* 16.6 (2015): 321.
- [8] Ganguli, Ishani, et al. "What Do High-Risk Patients Value? Perspectives on a Care Management Program." *Journal of general internal medicine* (2018): 1-8.

Plagiarism Report

Report Title:	Health Care Cost Analysis and Prediction
Report Link: (Use this link to send report to anyone)	https://www.check-plagiarism.com/plag-report/23592d427e2b17c3935b53aaebefe7c386af81615751870
Report Generated Date:	15 March, 2021
Total Words:	1493
Total Characters:	10441
Keywords/Total Words Ratio:	81.84%
Excluded URL:	No
Unique:	94%
Matched:	6%

Paper Work

HEALTH CARE COST ANALYSIS AND PRIDITION

¹Akshaya Dantha, ²Amrapali Bhowmik and ³P.Jeyanthi

^{1,2} UG Student, ³Professor, Dept. of IT, Sathyabama Institute of Science and Technology, Chennai.

¹akshayadantha@gmail.com, ²bhowmikamrapali111@gmail.com, ³jeyanthi.it@sathyabama.ac.in

Abstract- This project studies the temporal consistency of health care expenditures in a large state Medicaid program. Predictive machine learning models were used to forecast the expenditures, especially for the high-cost, high-need (HCHN) patients. We systematically tests temporal correlation of patient-level health care expenditures in both the short and long terms. The results suggest that medical expenditures are significantly correlated over multiple periods. Our work demonstrates a prevalent and strong temporal correlation and shows promise for predicting future health care expenditures using machine learning. Temporal correlation is stronger in HCHN patients and their expenditures can be better predicted. Including more past periods is beneficial for better predictive performance. This study shows that there is significant temporal correlation in health care expenditures. Machine learning models can help to accurately forecast the expenditures. These results could advance the field toward precise preventive care to lower overall health care costs and deliver care more efficiently.

Keywords: health management system, predictive analysis, LO

1.INTRODUCTION

Health care is one of the largest components of the global economy. According to the World Bank, in 2019, health care expenditures accounted for 9.95% of the world's total gross domestic product (GDP). This amount is expected to increase over the next several years, because of the expansion of insurance coverage under the Affordable Care Act. In addition, resources, in terms of expenditures, are disproportionately consumed by a relatively small proportion of the health care utilizing population. As a result, this group of health care utilizers has been termed high-cost, high-need (HCHN) patients due to its disproportionate spending concentration and highly prevalent comorbid chronic condition profile. Stakeholders have argued the need for higher efficiency health care, especially for HCHN patients. For example, managed care organizations and health plans (MCOs) have been deployed and funded under capitation payment systems to incentivize health care providers to deliver more cost-effective services.

Information technology provides a new, promising way to approach a wide-range of health care problems, especially in the "Big Data" era. Health care utilization routinely generates vast amounts of data from sources ranging from electronic medical records, insurance claims, vital signs, and patient-reported outcomes. Moreover, predicting health outcomes using data modeling approaches is an emerging field that can reveal important insights into factors associated with disproportionate spending patterns. Specifically, if we are able to forecast expenditures at the patient-level with good accuracy, we could improve targeted care by anticipating health care needs of HCHN patients. Moreover, predictive modeling can improve our understanding of causal pathways leading to our understanding of expensive events which informs system-level strategies for prevention. Indeed, prevention is one of the most effective ways to lower health care expenditures while delivering better quality of care.

HCHN patients have shown a high degree of persistence in medical expenditures reflected in administrative claims data, although some recent studies have concluded that the high utilization may be temporary and not persistent. The purpose of this study is to better understand these temporal patterns and to apply machine learning-based models to predict expenditures. Additionally, we would like to determine if practical timeframes (e.g. 1 month,

6 months) are feasible choices for predicting expenditures.

2. LITERATURE SURVEY

Healthcare Management System is achieved with the advent of Predictive Analytics. While base line of a Predictive Analytic Tool is the same, the tool comes in various versions. Several developers aim at developing a tool that is efficient than all other tools developed so far. The following papers have implemented some of the most prevalent tools.

2.1 DATA MINING

Data mining is defined as the analysis of often large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Data mining in the field of Healthcare is a gradually developing methodology that is used to identify a legitimate, useful relationships and patterns in the data that are too complex and subtle for human to make out. The previously unknown patterns and relationships in a database can be easily identified with a Data mining algorithms like classification, association, clustering, etc and with this we can build a

Predictive Model. To identify the unknown patterns, we use the life cycle model of data mining. It includes all the process from data selection to data exploitation.

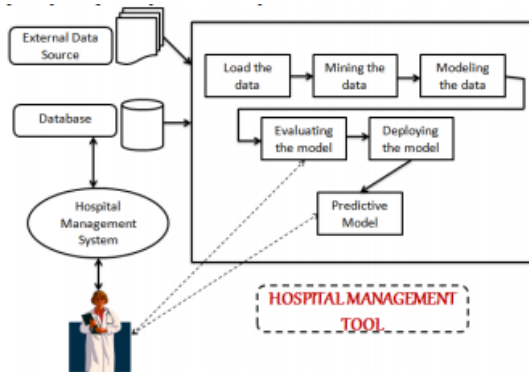
2.2 DATA MINING IN HEALTH CARE

Data mining is beneficial for all those who are involved in the healthcare including Healthcare Insurers, Healthcare organizations, Doctors and patients. For healthcare insurers, data mining can help identify Fraud and Abuse in the insurance. For Healthcare organizations, it supports Hospital Management system, Patient Tracking system and thereby forming a strong Customer relationship management. For doctors and physicians, it can help predict the LOS, mortality rate, risk factors, and re-admission rate. With all the aforesaid, a doctor can choose upon the effective treatment available for the patient. And finally, a patient can get a quality treatment at affordable cost. According to a survey, 87% of the deaths in the US hospital have been prevented with the Predictive tool.

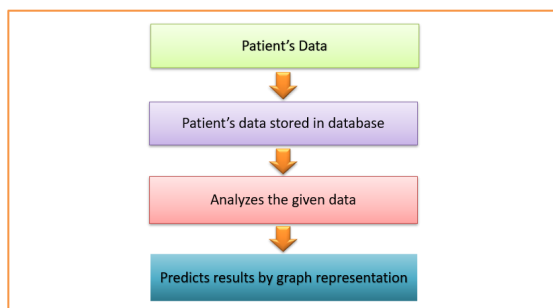
2.3 ADVANTAGES OF DATA MINING IN HEALTH CARE

With the advent of Information Technology in the field of healthcare, the paper based prescription has turned to Electronic records, which is highly reliable. The electronic record is capable of storing voluminous data and this includes patient demographic, socio-economic data, previous medical data, laboratory values and

much more. These data though private are essential for providing an effective treatment to a patient. Data mining helps in providing a clear policy for privacy and security.



3.SYSTEM ARCHITECTURE



4.METHADODOLOGY

Supervised learning problems can be further grouped into

- Regression problems and
- Classification problems

In classification, learning algorithms takes the input data and map the output to a discrete output like True or False In regression, learning algorithms maps the input data to continuous output like weight, cost, etc.

In this project I will apply regression techniques of supervised learning to predict the medical insurance costs.

Dataset consists of records. Each record contains the below data for specific person.

- age – Age of the person
- sex – Sex of the person
- bmi – Body Mass Index(BMI) of the person
- children – Number of children for the person
- smoker – Smoking status of the person

- region – Region of the person in US
- charges – Medical Insurance costs per year for the person

To understand correlation between the input or the independent variables, I classified the ages and BMI into specific categories. The above analysis shows the stats like number of male and female, smokers and non-smokers, etc. from the input data. Taking the BMI ranges from Medline Plus, I categorized the input samples into below-weight, normal-weight, over-weight and obese.

3. Splitting the data set into training and testing subsets helps to assess the performance of the model over an independent data set. Typically, we train the model using training data subset and then evaluate the model's performance using the testing data subset, which is independent of the training data subset.
4. Splitting the data set also helps in having a check on model's overfitting.

As part of evaluation, I considered multiple regression algorithms like decision trees, Support Vector Machines for regression, etc. Based on the metrics, I choose to use decision tree technique for this project

The goal of decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

5.RESULTS

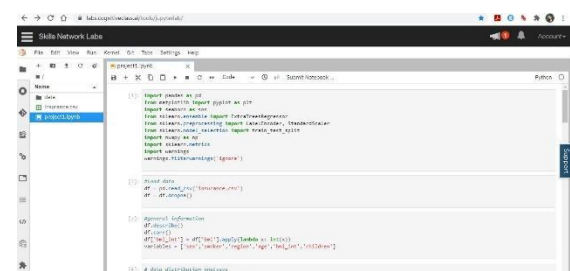


Figure 1: code

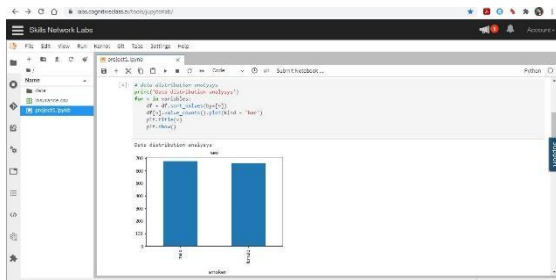


Figure 2: Data Distribution Analysis of Sex

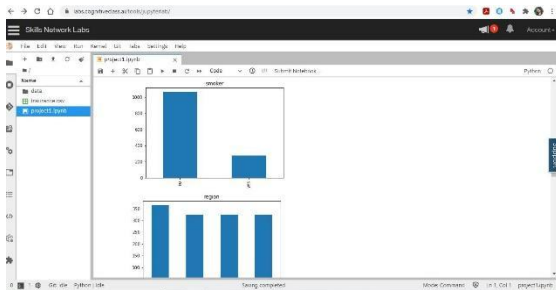


Figure 3: Data Distribution Analysis of Smoker and Region

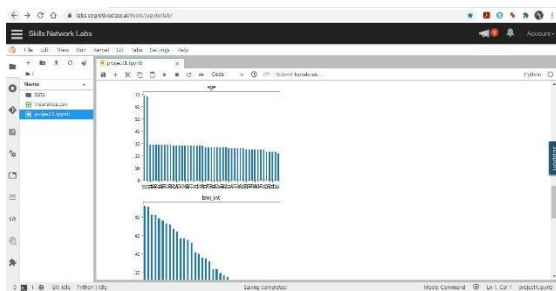


Figure 4: Data Distribution Analysis of age

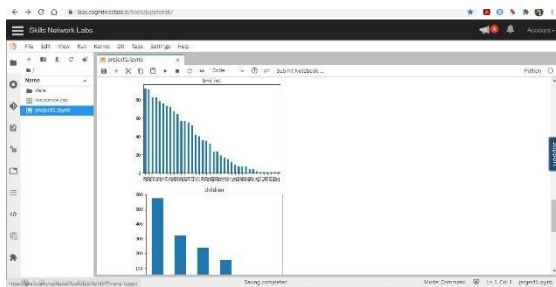


Figure 5: Data Distribution Analysis of bmi_int

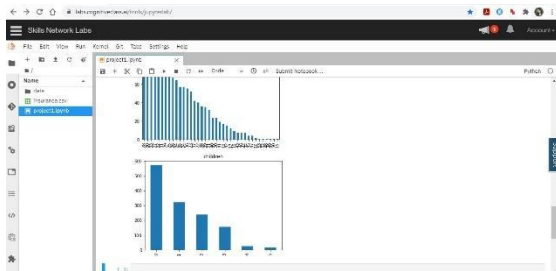


Figure 6: Data Distribution Analysis of Children

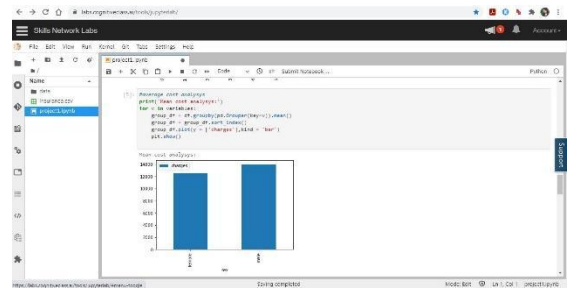


Figure 7: Mean Cost Analysis of Sex

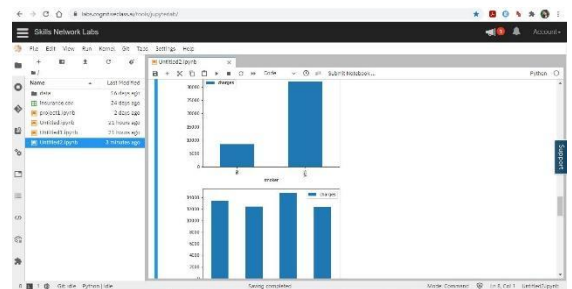


Figure 8: Mean Cost Analysis of Smoker and Region

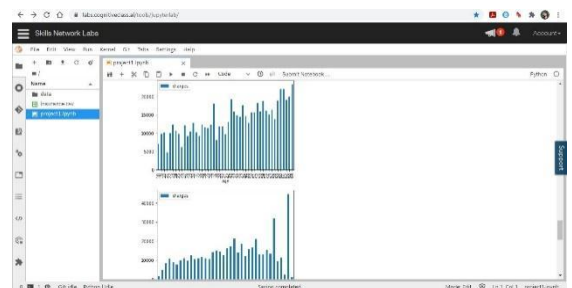


Figure 9: Mean Cost Analysis of Age and bmi_int

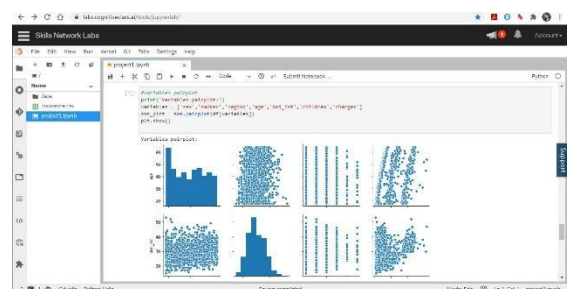


Figure 10: Variables pairplot



Figure 11: Variables pairplot

6.CONCLUSION

We systematically tests temporal correlation of patient-level health care expenditures in both the short and long terms. The results suggest that medical expenditures are significantly correlated over multiple periods. Our work demonstrates a prevalent and strong temporal correlation and shows promise for predicting future health care expenditures using machine learning. Temporal correlation is stronger in HCHN patients and their expenditures can be better predicted. Including more past periods is beneficial for better predictive performance.

7.REFERENCES

[1] Precision Medicine Initiative (NIH). <https://www.nih.gov/precision-medicine-initiative-cohort-program>.

[2] Lyman, Gary H., and Harold L. Moses. "Biomarker Tests for Molecularly Targeted Therapies--the Key to Unlocking Precision Medicine." *The New England journal of medicine* 375.1 (2016): 4-6.

[3] Collins, Francis S., and Harold Varmus. "A new initiative on precision medicine." *New England Journal of Medicine* 372.9 (2015): 793-795.

[4] Xu, Rong, Li Li, and QuanQiu Wang. "dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text." *BMC bioinformatics* 15.1 (2014): 105.

[5] Wang, Bo, et al. "Similarity network fusion for aggregating data types on genomic scale." *Nature methods* 11.3 (2014): 333.

[6] Tatonetti, Nicholas P., et al. "Data-driven prediction of drug effects and interactions." *Science translational medicine* 4.125 (2012): 125ra31-125ra31.

[7] Libbrecht, Maxwell W., and William Stafford Noble. "Machine learning applications in genetics and genomics." *Nature Reviews Genetics* 16.6 (2015): 321.

[8] Ganguli, Ishani, et al. "What Do High-Risk Patients Value? Perspectives on a Care Management Program." *Journal of general internal medicine* (2018): 1-8.

Source Code

```
import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

from sklearn.ensemble import ExtraTreesRegressor

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.model_selection import train_test_split

import numpy as np

import sklearn.metrics

import warnings

warnings.filterwarnings('ignore')

#load data

df = pd.read_csv('insurance.csv')

df = df.dropna()

#general information

df.describe()

df.corr()

df['bmi_int'] = df['bmi'].apply(lambda x: int(x))

variables = ['sex','smoker','region','age','bmi_int','children']

# data distribution analysys

print('Data distribution analysys')

for v in variables:
```

```
df = df.sort_values(by=[v])
df[v].value_counts().plot(kind = 'bar')
plt.title(v)
plt.show()
```

```
#average cost analysys
```

```
print('Mean cost analysys:')

```

```
for v in variables:
```

```
    group_df = df.groupby(pd.Grouper(key=v)).mean()
    group_df = group_df.sort_index()
    group_df.plot(y = ['charges'],kind = 'bar')
    plt.show()
```

```
#variables pairplot
```

```
print('Variables pairplot:')

```

```
variables = ['sex','smoker','region','age','bmi_int','children','charges']

```

```
sns_plot = sns.pairplot(df[variables])

```

```
plt.show()
```

```
print('Model training and evaluating\n\n')
```

```
#transform categorical data
```

```
le_sex = LabelEncoder()
```

```
le_smoker = LabelEncoder()
```

```
le_region = LabelEncoder()
```

```
df['sex'] = le_sex.fit_transform(df['sex'])
```

```
df['smoker'] = le_smoker.fit_transform(df['smoker'])
```

```

df['region'] = le_region.fit_transform(df['region'])

variables = ['sex','smoker','region','age','bmi','children']

X = df[variables]
sc = StandardScaler()
X = sc.fit_transform(X)
Y = df['charges']
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)

#train model
regressor = ExtraTreesRegressor(n_estimators = 200)
regressor.fit(X_train,y_train)

#prediction and evaluation
y_train_pred = regressor.predict(X_train)
y_test_pred = regressor.predict(X_test)

print('ExtraTreesRegressor evaluating result:')
print("Train MAE: ", sklearn.metrics.mean_absolute_error(y_train, y_train_pred))
print("Train RMSE: ", np.sqrt(sklearn.metrics.mean_squared_error(y_train,
y_train_pred)))
print("Test MAE: ", sklearn.metrics.mean_absolute_error(y_test, y_test_pred))
print("Test RMSE: ", np.sqrt(sklearn.metrics.mean_squared_error(y_test,
y_test_pred)))

print('Feature importance ranking\n\n')

```

```

importances = regressor.feature_importances_

std = np.std([tree.feature_importances_ for tree in regressor.estimators_],axis=0)

indices = np.argsort(importances)[::-1]

importance_list = []

for f in range(X.shape[1]):

    variable = variables[indices[f]]

    importance_list.append(variable)

    print("%d.%s(%f)" % (f + 1, variable, importances[indices[f]]))


# Plot the feature importances of the forest

plt.figure()

plt.title("Feature importances")

plt.bar(importance_list, importances[indices],

        color="r", yerr=std[indices], align="center")

plt.show()


print('Predicting on new data\n\n')


billy = ['male','yes','southeast',25,30.5,2]

print('Billy - ',str(billy))


billy[0] = le_sex.transform([billy[0]])[0]

billy[1] = le_smoker.transform([billy[1]])[0]


billy[2] = le_region.transform([billy[2]])[0]

```

```
X = sc.transform([billy])
```

```
cost_for_billy = regressor.predict(X)[0]
```

```
print('Cost for Billy = ',cost_for_billy,'\n\n')
```

```
dennis = ['female','no','southeast',45,19,0]
```

```
print('Dennis - ',str(dennis))
```

```
dennis[0] = le_sex.transform([dennis[0]])[0]
```

```
dennis[1] = le_smoker.transform([dennis[1]])[0]
```

```
dennis[2] = le_region.transform([dennis[2]])[0]
```

```
X = sc.transform([dennis])
```

```
cost_for_dennis = regressor.predict(X)[0]
```

```
print('Cost for Dennis = ',cost_for_dennis)
```